DOCUMENT RESUME

ED 399 294 TM 025 609

AUTHOR Johanson, George A.
TITLE A Compromise Grading Model for Classroom Tests.
PUB DATE Apr 92
NOTE 6p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Cutting Scores; *Educational Testing; Error of Measurement; *Grading; *Norm Referenced Tests; Sampling; Secondary Education; *Secondary School Students; Standards; *Test Construction; Test Use
IDENTIFIERS *Compromise

ABSTRACT

Most educational measurement texts distinguish between norm-referenced (NR), or relative, methods of assigning letter grades to objective test scores, and criterion-referenced (CR), or absolute, methods. Both NR and CR approaches have serious limitations in typical classroom situations, and neither approach, in its pure form, may be entirely suitable. An alternative method is proposed and illustrated with scores from 57 secondary school students taking a 26-item objectively scored test. The approach involved using a smoothed or fitted cumulative distribution and a ratio of standard errors to fix the slope of the line through the ideal cut-points. This is a modification of the method of C. H. Beuk (1984). The rationale for this type of compromise is that it acknowledges the sample status of both the set of test items and the group of examinees and shares sampling error equally between NR and CR methods. The algorithm has been programmed in PASCAL for the microcomputer. A structured grading method of this sort would allow teachers of multiple sections or those within the same department to give somewhat comparable grades to their students if they used agreed-on NR standards and individual CR standards. This compromise would be especially useful when an entirely new test is used or an unfamiliar group of students is encountered. (Contains 1 table, 2 figures, and 10 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
GEORGE JOHANSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 399 294

A COMPROMISE GRADING MODEL FOR CLASSROOM TESTS

George A. Johanson, College of Education, Ohio University

A Paper Presented at the American Education Research Association Annual Meetin
San Francisco, April 1992

Objectives

Methods for assigning letter grades to a set of objective test scores would seem to be a somewhat neglected area of technical concern. Most educational measurement texts distinguish between norm-referenced (NR) or relative methods and domain/criterion- referenced (CR) or absolute methods. However, both NR and CR approaches can be seen to have serious limitations for use in typical classroom situations. In a rather comprehensive examination, at a relatively low cognitive level (for example, in mastery learning), and with relatively few students, absolute standards such as percentage-correct seem more appropriate in the sense of yielding interpretable scores (e.g., percent mastery of the content domain). If there are many students, testing is at a higher cognitive level, and the test is comprised of a less than comprehensive set of items, then relative standards (such as z-scores) might seem more interpretable (e.g., percentile rank in the population). That is, if the sample of students is large enough to be representative of the population and to yield accurate (percentile) estimates of each student's relative position in the population, then NR may be a viable approach to grading. Conversely, if the content domain is clearly defined (factual) as opposed to implied (higher level skills) and the sample of items is large enough to accurately reflect the content domain and to yield accurate estimates of each student's score (proportion of items correct), then CR may be an appropriate grading method.

In a usual classroom situation, there might be 20-60 examinees with 20-60 test items at a variety of cognitive levels and therefore neither approach, in its pure form, seems particularly well-suited. Recognizing these and other factors, Terwilliger (1989) recommends using CR for some grading decisions and NR for others.

A compromise between NR and CR seems both reasonable and consistent with current practice. Consistent, in that many teachers use absolute standards in the form of percent-correct (sometimes because of school or district policy), but then 'adjust' the raw scores in a variety of ways if the distribution of grades seems inappropriate or improbable. Indeed, some teachers relying solely on NR grading are quick to admit that they also have CR 'limits' and will, for example, not award an 'A' to any score below a certain percentage-of-items-correct. A compromise is reasonable since a valid interpretation of a CR or a NR grade requires knowledge of either the content domain or the population of students, respectively. The blending of the two approaches might better reflect the actual partial knowledge of both the content domain and population by the typical consumer of the grade. Indeed, grades are often seen to reflect, to some extent, both absolute and relative achievement.

From the responses of students in my measurement classes, the most popular procedures for adjusting the scores would seem to be 'gapping' or 'eyeballing' (Wainer & Schacht, 1978), adding a fixed number (or percentage) of points to everyone's score, dropping items, or simply making sure the next test results in grades with a compensating distribution. It is somewhat ironic that if both the mean and standard deviation of a domain-referenced test are adjusted using a linear transformation then we have the equivalent of a most prevalent form of norm-referencing, the z-score. The focus of this paper is on an alternative method of adjustment.

Theoretical Framework and Example with Real Data

Hofstee (1983) has suggested using a cumulative frequency distribution to better see the relationship between NR and CR decision-making. While Hofstee did this in reference to large-scale testing, some of the principles involved apply equally well to the classroom situation. Figure 1 is an example of a cumulative number- correct frequency distribution. The scores are from 57 secondary students taking a 26-item objectively-scored test. A score of 15 would be approximately at the 45th percentile. Since there were 5 grading categories: A, B, C, D, F, we can identify the expected outcomes by locating points 1-4 using both the absolute standards we have set and our past grading practice with this unit of study. That is, if we have observed, over many sections, that 16% of the students received A's, 23% received B's, 18% received C's, 20% received D's, and 23% received F's, then the NR standards may be seen on the vertical axis where the cumulative proportion of students below each number-correct score is given. If it is felt that 93% or more of the items must be answered correctly to receive an A, 85% or more to receive a B, 75% or more to receive a C, and 65% or more to receive a D, then the corresponding number-correct standards can be seen on the horizontal axis. The intersections of the expected cut-points (1-4) are such that both the NR and CR expectations are simultaneously met if and only if the observed distribution passes through these points. That is, if these intersections are on the observed cumulative distribution, then we are done; if not, some decision or compromise is necessary. In a sense, the intersections are points on our best estimate of a population cumulative distribution.

The particular compromise suggested by Hofstee involves setting minimum and maximum acceptable percentages-correct about each expected CR cut-point and minimum and maximum acceptable proportions of students in each grading category about each NR cut-point to determine the slope of a diagonal line through the meeting point. This line is then extended to meet the observed cumulative distribution and the

TM 025 609

BEST COPY AVAILABLE

intersection is the compromise.

Beuk (1984) proposed that the compromise be obtained by using the ratio of standard deviations of the ratings of a group of judges as the slope of the diagonal. De Gruijter (1985) uses estimates of the uncertainties concerning both the NR and CR ideals to define a family of ellipses, selects the tangent ellipse, and uses the abscissa of the intersection as the compromise cut-score.

These procedures all require additional judgments and do not directly take into consideration the notion that, all else being equal, with larger numbers of examinees and fewer items it might be reasonable to depend more heavily upon the NR criteria and vice-versa. That is, if we have a primarily CR grading philosophy, then we would be concerned that there are a sufficient number of items to adequately represent a clearly identified content domain and to permit accurate estimation of a student's domain score. If primarily NR, the concern would be to have a sufficient number of students (representative of the population) in the sample to accurately estimate a student's percentile rank. In a compromise situation, more weight might be given to the more accurate estimation at each decision level.

An additional problem is encountered using the observed cumulative distribution. As the ratio of test length to number of students increases, there will be more and larger gaps or zero frequencies in the frequency distribution and these will be seen as 'flat spots' in the cumulative distribution. Due to these random gaps and other sample fluctuations, the need to smooth the observed distribution arises. While there are a number of smoothing approaches available, the beta-binomial (or negative hypergeometric) model has been found to be a most efficient presmoother for equipercentile equating (Fairbank, 1987) and has also been successfully used to model number-correct achievement test score data (Duncan, 1974; Keats & Lord, 1962). Lord and Novick (1968) recommend this model for fitting observed distributions of number-correct scores and provide a theoretical rationale for the model. A convenient algorithm for computing the beta-binomial is available (Huynh, 1979).

Method

The approach followed was to use a smoothed or fitted cumulative distribution and to use a ratio of standard errors to fix the slope of the line through the ideal cut-points. This is a modification of Beuk's method in that his ratio of standard deviations is also a ratio of standard errors (the same judges are used to provide both standard deviations). It is important to note, however, that there is no variation in our ideal cut-points; these points may be thought of as being akin to population parameters.

If we conceptually fix a CR standard and the test, then each sample of students or class from our (assumed infinite) population of students will yield a sample proportion of students at or below this CR standard and this sample proportion may be compared to the hypothesized (population) proportion. The standard error of such proportions is given by $(\pi_n(1-\pi_n)/n)^{0.5}$ where n is the number of examinees and π_n is the population proportion or CR standard. In exactly the same way, we may imagine a single NR standard (proportion of items correct=π_k) and class or group of students as fixed and compute the standard error of the proportion of items answered correctly, $(\pi_k(1-\pi_k)/k)^{0.5}$, as if the k items on our test were a sample from the (assumed infinite) content domain. The compromise is to use the ratio of these standard errors, $[(\pi_n(1-\pi_n)/n)^{0.5}]/[(\pi_k(1-\pi_k)/k)^{0.5}]$, as the slope of the line through the ideal points.

The rationale for this type of compromise is that it acknowledges the sample status of both the set of test items and the group of examinees and shares the sampling error equally between methods (NR and CR). In particular, the sample of items is given the same credibility as the sample of students in that the compromise at each decision level departs from the NR and CR standards by the same number of standard errors.

In practice, standard errors are largely influenced by sample size and this means that when the ratio of number of items to number of examinees is large, the tendency will be for the compromise to rely more heavily on the CR standards. When the ratio is small, the compromise will rely more heavily on the NR standards. That is, reliance is placed on both NR and CR standards, but the compromise at each decision point tends to proportionally favor the standard with the smaller standard error.

In the example, the CR cut-point between a grade of B and C was to be a percentage-correct score of 85%. The standard error of a proportion of items is $(\pi_k(1-\pi_k)/k)^{0.5}$ where k is the number of items on the test or, $(0.85*(1-0.85)/26)^{0.5} = 0.070$. The corresponding standard error of a proportion of persons below a grade of B is $(\pi_n(1-\pi_n)/n)^{0.5}$ where n is the number of examinees or, $(0.39*(1-0.39)/57)^{0.5} = 0.065$. The resulting ratio of 0.065/0.070 = 0.923 would be negated, converted to the number-correct scale, and used as the slope of the line through point 3, see Figure 2. Linear interpolation is then used with the smoothed cumulative distribution and the resulting abscissa of the point of intersection (b) is 18.047. This is the suggested compromise B/C cut-score shown in Figure 2 for grading purposes with this test given the CR and NR parameters. The calculations for the other three cut-points are similar. Note that the slopes are all less than 1 for this example. This is the result of somewhat greater reliance on the NR standards than on the CR standards in arriving at the compromise since there were 57 students and 26 test items. The ratio of standard errors, however, is not just the ratio of number of persons to number of items, but also reflects the expected proportions.

By using a constant times the ratio of standard errors, we could adjust these cut-scores to yield any desired weighting of NR-CR

3 BEST COPY AVAILABLE

standards, perhaps to better reflect the cognitive level of the majority of the test items. It is interesting to note that setting the constant (and hence the slope) to a value near zero results in an equipercentile equating of the smoothed observed score distribution to the 'norming group' distribution defined by the ideal points. This would seem to be a reasoned method of relative grading if the ideal points were derived from data over many sections.

To make the scores more understandable to students and others, it may be desirable to follow the popular practice of presenting the results as adjusted raw scores or adjusted percentages that can then be compared to the stated standards for letter grade decisions. The scores, percentages, adjusted scores, and adjusted percentages are shown in Table 1. Letter grades for this example are also shown in Table 1 where the NR letter grades were calculated using z-scores with cut-scores that reflect the expected percentages of A's, B's, and so on. The suggested or compromise letter grades are in the last column labeled NR/CR. Note how the NR/CR grades mediate the NR and CR grades somewhat differently at each score level. This procedure is not equivalent to a simple 'averaging' of NR and CR grades.

For the example data, the mean number-correct score is 15.47, the standard deviation is 4.16, and the reliability (KR-21) is 0.66. Using a Kolmogorov-Smirnov one sample test of fit, the maximum absolute difference is 0.095. The null hypothesis (model fits the data) is accepted at $p = 0.985$. The beta-binomial has successfully fit (conservatively, at $\alpha = 0.20$) over 95% of real data sets so far investigated and has fit 100% of the author's classroom data for the past two years.

The algorithm has been programmed in (standard) Pascal for the IBM microcomputer. There are several additional outputs, the program can be run in batch mode or interactively, and there is an accompanying document. It is available without cost from the author when the request is accompanied by a formatted disk and stamped mailer.

Conclusions and Educational Importance

Grades are important: they are the coin-of-the-realm in education. Many teachers find the task of evaluation difficult and might welcome a structured method for obtaining, at least, suggested letter grades in those situations where adherence to absolute standards would result in an unacceptable distribution of letter grades.

Continuing to adjust proportion-correct standards by *ad hoc* methods is neither reasoned nor reliable. Structured grading methods such as this would also permit teachers of multiple sections or those within the same department to give somewhat comparable grades to their students if they used agreed-upon NR standards and individual CR standards that reflect professional judgement about differences in the difficulties and objectives of their individual tests. The use of a computer to assist in grading decisions means that practical and useable approaches need not be overly simplistic. This compromise might prove most useful when an entirely new test is used or when an unfamiliar group of students is encountered. When a teacher is obliged to adhere to grading standards as in the example (93% and above for an 'A'), giving tests that challenge all students and that reflect higher level cognitive skills becomes virtually impossible without some means of score adjustment. 'Eyeballing' a set of scores is simply not good grading practice.

References

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21, 147-152.

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. Journal of Educational Measurement, 22, 263-269.

Duncan, G. T. (1974). An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 69, 50-57.

Fairbank, B. A. (1987). The use of presmoothing and postsmoothing to increase the precision of equipercentile equating. Applied Psychological Measurement, 11, 245-262.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson, & J. S. Helmick (Eds.), On educational testing (pp. 109-127). San Francisco: Jossey-Bass.

BEST COPY AVAILABLE

Huynh, H. (1979). Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics, 4, 231-246.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika, 27, 59-72.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Terwilliger, F. S. (1989). Classroom standard setting and grading practice. Educational Measurement: Issues and Practice, 8, 15-19.

Wainer, H., & Schacht, S. (1978). Gapping. Psychometrika, 43, 203-212.

Figure 1 Cumulative distribution of the observed number-correct raw scores.

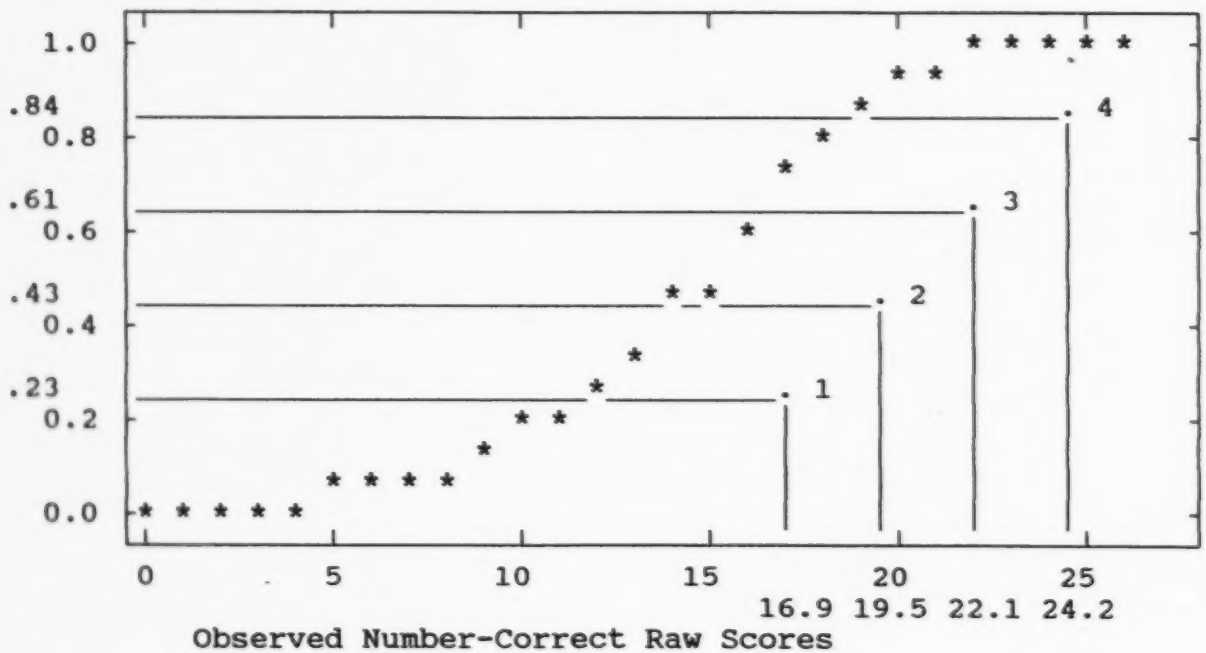

Figure 2 Cumulative distribution of the smoothed number-correct raw scores.

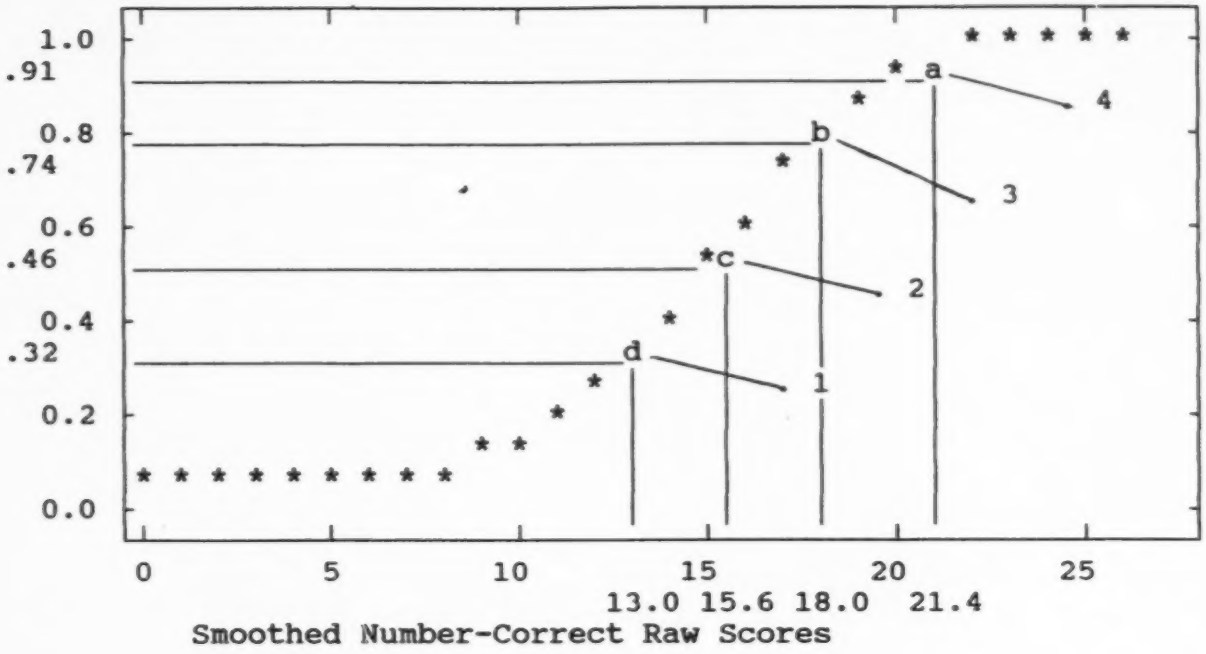

Table 1 Example Data With NR, CR, and Compromise Grades

X	Freq	%	Adj. X	Adj. %	z-score	NR	CR	NR/CR
23	2	88.5	24.81	95.44	1.808	A	B	A
22	3	84.6	24.42	93.91	1.568	A	C	A
21	2	80.8	23.93	92.05	1.328	A	C	B
20	2	76.9	23.31	89.66	1.087	A	C	B
19	6	73.1	22.69	87.27	0.847	B	D	B
18	2	69.2	22.05	84.80	0.607	B	D	C
17	8	65.4	20.97	80.67	0.367	B	D	C
16	6	61.5	19.90	76.53	0.126	C	F	C
15	3	57.7	18.87	72.57	-0.114	C	F	D
14	5	53.8	17.87	68.72	-0.354	D	F	D
13	5	50.0	16.86	64.84	-0.594	D	F	F
12	2	46.2	15.56	59.85	-0.835	F	F	F
11	2	42.3	14.26	54.86	-1.075	F	F	F
10	5	38.5	12.97	49.87	-1.315	F	F	F
9	1	34.6	11.67	44.89	-1.555	F	F	F
8	2	30.8	10.37	39.90	-1.796	F	F	F
5	1	19.2	06.48	24.94	-2.516	F	F	F

Note. X is the raw number-correct score; NR/CR is the compromise grade.